ALGORITHM

In customer segmentation, different algorithms and techniques have been studied and implemented throughout the years.

First of all, it is important to note that, while clustering is the most common technique[reference], it is not the only one applied in this field, as some studies have used predictive techniques (i.e., use machine learning methods to generate a model that can predict an outcome, based on a set of input). Some predictive algorithms used can be as simple as decision trees or more complicated supervised classification methods, such as Neural Networks and Support Vector Machines.[2008 sc] Although classification may be useful in segmenting the customers, this requires additional pre-processing and, as they are supervised models, need a huge amount of data and appropriate algorithms to work properly; also, the definition of the target variable (the one that needs to be predicted) is not trivial.

**Clustering** thus remains the most favourable technique. However, different types of clustering algorithms exists and the correct choice of it surely makes the difference in the result. In general, there are five categories of clustering algorithms:

1. Centroid-based: data is organized into groups, where a datapoint belongs to the group that has the centre closest to the point. Centroid-based algorithms are efficient but sensitive to initial conditions and outliers, and often requires the specification of the number of cluster a-priori.
2. Density-based: connect areas of high density into clusters, thus permitting arbitrary-shaped distributions. These algorithms have difficulty with data of varying densities and high number of dimensions but are robust to outliers.
3. Distribution-based: these assume data is composed of distributions, and the probability of a point belonging to a distribution depends on the distance to the distribution’s centre. It can only be used when the distribution of the data is known.
4. Hierarchical: these algorithms create trees of clusters. They are particularly suited for data where a hierarchical structure is clearly identifiable.
5. Neural networks: use networks to perform clustering.

Centroid-based

The majority of papers focus on the **k-means** algorithm for customer segmentation [reference], that is a fast centroid-based algorithm. Multiple modifications have been proposed to handle the main downsides of this algorithm:

* Initialization problem:
* Determining the number of clusters:
* Handling noise:
* Hybrid alg:

Still, also the basic version of k-means has been used for customers and geographic segmentation [].

Another algorithm widely used for both geographical and customer segmentation is the **fuzzy c-means** algorithm: it is a centroid-based algorithm too but gives the possibility to have data points that belong to more than one cluster(overlapping), at the cost of higher required time. It is found that in some markets, for instance in the e-commerce, “soft clustering produces more promising results for real-world applications than hard clustering, through which each customer can be assigned to only one cluster.”[2011]

Density-based

**DBSCAN**, a density-based algorithm, has also been used for customer segmentation, and its performance have been proved similar of the one from k-means[2017 cs]. Improvement for DBSCAN have been proposed, such as NQ-DBSCANS and ρ-Approximate DBSCAN, but none have been applied to customer segmentation yet.

Distribution-based

Distribution based clustering is not applicable in the case studied, as no distribution is available or easily estimated.

Hierarchical

For segmentation problems, research suggests that hierarchical approaches do not perform very well with large data sets[Kuo et al., 2002a], as hierarchical methods build a tree structure using a dendogram that above a certain level becomes imprecise.

NN

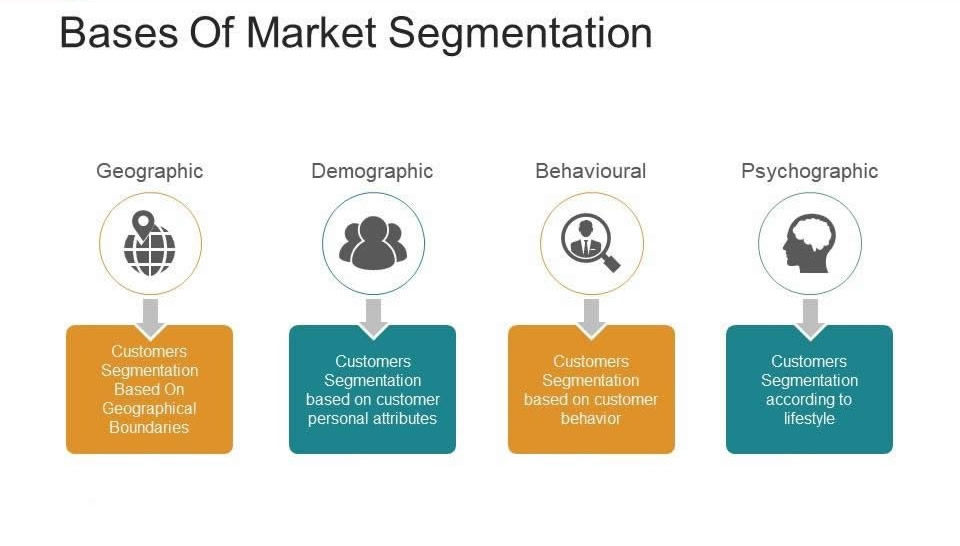
Finally, regarding neural network-based clustering algorithms, the most popular technique is self-organizing map (**SOM**), also called Kohonen map. It has been used for successfully segment customers in multimedia and telecommunication markets.[][] This technique can also be used for data pre-processing before applying another algorithm (usually k-means).

Others

Some improved algorithms, such as consensus clustering (the usage of more than one cluster algorithm in parallel or in series) or **two-phase clustering** (the usage of one cluster algorithm in combination with another) have been considered but are seldomly used for customer segmentation [2013]. Among the algorithms discussed above, the best one is dependant of the data is applied to and cannot be decided a priori.

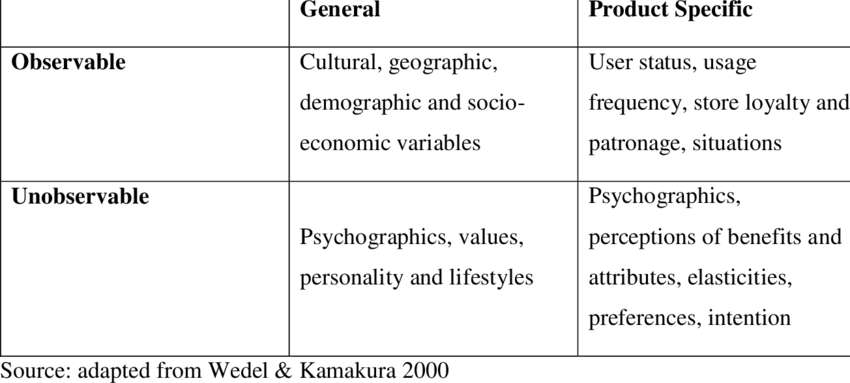
SEGMENTATION BASES

As important as the choice of the algorithm is the choice of the variables used: different ways of segmenting customers exist based on the segmentation variables (or bases) used. The most common bases are geographic, demographic, behavioral, and psychographic [2013], but multiple others are present such as distribution, media, price, time and occasion-based. [2021]



Variables can also be categorized as: [2013].

* General observable variables (e.g., geographic, demographics, socio-economic variables)
* Product specific observable variables (e.g., usage frequency and loyalty)
* General unobservable variables (e.g., lifestyle, psychographics)
* Product specific unobservable variables (e.g., benefits, preferences and intentions)



While observable variables are the cheapest and easiest one to collect, the information that can be retrieved by their analysis is not always enough, depending on the market analyzed and the insights that are expected: the sole usage of demographic variable has been proven ineffective in certain markets like FMCG and retail, while some extended demographic variables like education, house size and income have been proved effective. [2011 review] The selection of basis of market segmentation is completely dependent upon the industry and product type itself, and it can be stated that demographic may prove as good basis for segmentation.[] It is important to remember that, for segmentation in support of the launch of a new product, demographics are not enough while non-demographic traits such as values, tastes and preferences are more likely to influence consumers’ purchasing[1964],but this is not in the purpose of the project.

Demographic variables, as said, are one of the most used bases due to its low obtaining cost. Geographic variables are often used in combination with them, following the idea that “people that live up close have more similar characteristics”: these variable, by themselves, have been proved misleading and dangerous, as differences exist among customers who are in the same geographic location. It is thus always advised to use them with demographics, providing an additional option to locate the customers. [2009] RFM (recency, frequency, monetary value) and LTV (customer life-time value) are also between the most common input variable used in the literature for clustering customers, [2018] and belongs to the behavioural class of bases. RFM, in particular, is very popular as it uses fewer segmentation variables [2021], making it simple, easy to implement, and easy to understand by managers and decision-makers. Following the Pareto principle of “80% of sales come from 20% of clients”, Clustering based on RFM can help label groups of customers from “golden” to “dormant”, [2021 forse] taking into account their prior purchasing history. RFM analysis can also take in consideration the impact of a product in the market, as it makes easier to grasp changes in the clusters when new products are release. Different bases serve different purpose and choosing the correct base can make the difference in soundness of segmentation.

EVALUATION

standarfization/normalization, reliability and validity